WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

Identifying Rank-Influential Groups of Observations in Linear Regression Modeling

Peter J. Kempthorne

Sloan Working Paper # 3018-89-MS

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

Identifying Rank-Influential Groups of Observations in Linear Regression Modeling

Peter J. Kempthorne

Sloan Working Paper # 3018-89-MS

Identifying Rank-Influential Groups of Observations in Linear Regression Modeling

Peter J. Kempthorne*
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

May 23, 1989

Abstract

When fitting a linear regression model, deleting a group of observations from the data may decrease the precision of parameter estimates substantially or, worse, render some parameters inestimable. These effects occur when the deletion of the group degrades the rank of the model matrix for the regression. We present theory and methods for identifying such rank-influential groups of observations.

Key words: Rank Deficiency; Influence; Linear Models; Projection Pursuit; Rank; Regression; Factor-analytic rotations

*This work was supported in part by Grants SES 8401422 and DMS-8601467 from the National Science Foundation. The author is grateful to S. Weisberg and participants of the 1989 Diagnostics Quarter at the University of Minnesota for their comments and discussion.

1 Introduction

Let (y, X) denote a data set where X is an $n \times p$ matrix of values of p explanatory variables for each of n cases and y is an $n-$vector of values for a single response variable for each case. When fitting the linear regression model $y = X\beta + \epsilon$, where β is the $p-$vector of regression parameters and ϵ is an $n-$vector of regression errors, the assumption that X has full rank is necessary for the estimability of every component of β. Even if X hs full rank, it may contain a group of cases, I, whose exclusion from the data would render X rank deficient. The presence of such a group in the data determines whether the linear model parameters are all estimable. We shall call such groups "rank influential."

Identifying rank-influential groups is important both before collecting the response variable data and afterward, during the analysis. Suppose a full-rank model matrix X is proposed, but a case group I exists for which the reduced matrix $X_{(I)}$ is rank deficient. If there is a nonnegligible probability that the response variable will be missing for all cases in I, then one might prefer to add more cases to the design to protect against the possibility that model parameters are inestimable. When analyzing the fit of a linear model, it could be important to know that a group of cases is strongly influential in that its absence from the data would render certain parameter estimates unavailable. Additional data collection and analysis might be appropriate because inferences based on these estimates depend critically on data from cases in the rank-influential group.

The importance of rank-influential case groups has been acknowledged in the literature; see, for example, Cook and Weisberg (1980, 1982) and Belsley, Kuh, and Welsch (1980). However, no practical method has been proposed for identifying such case groups except in the special case of groups of size one.

In Section 2, we address the issue of estimability and rank of the model matrix when fitting a linear regression model. We present mathematical characterizations of rank-influential case groups in Section 3. Section 4 proposes two methods for identifying such groups. These methods rely on factor-analytic rotation algorithms and projection pursuit techniques and are computationally feasible alternatives to direct combinatorial search algorithms. In addition to identifying rank-influential groups, the methods also identify case groups which

are potentially very influential on estimates of model parameters. Section 4 also details how the methods perform on both simulated and real data. Section 5 concludes with some discussion of related diagnostic methods.

2 Estimability and Rank Deficiency in Linear Models

Suppose the dependence of y on X follows a linear model:

$$y = X\beta + \epsilon \tag{1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of fixed unknown regression parameters and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is the vector of random errors with $E(\epsilon) = 0$ and $Cov(\epsilon) = E(\epsilon\epsilon^T) = \sigma^2 I_n$, where I_n is the order-n identity matrix and σ^2 is the unknown, positive variance of the errors.

It is well known that the regression parameter β is estimable only if X has full (column) rank. If X does not have full rank, then for any $\beta \in \Re^P$ specifying the linear model (1), there exist (infinitely many) $\beta' \neq \beta$ such that $X\beta = X\beta'$.

One method of determining whether a matrix X has full rank is to calculate $|X^T X|$, the determinant of the $p \times p$ matrix $X^T X$. The determinant is non-zero if and only if X has full (column) rank $p \leq n$. To interpret magnitudes of the determinant, note that when the distribution of ϵ is multivariate normal, the volume of a confidence ellipsoid for β is proportional to $|X^T X|^{-\frac{1}{2}}$. The smaller the determinant $|X^T X|$, the larger the confidence ellipsoid of a given level for β, and in the limit when X has deficient rank, the volume of any confidence ellipsoid is infinite.

Another method for assessing the rank of a matrix X is based on the singular value decomposition (SVD) of X. Golub (1969), Belsley, Kuh, and Welsch (1980), and Stewart (1984) discuss the SVD. The rank of a matrix X is the number of non-zero singular values in the SVD of X. So X has full rank if and only if its smallest singular value is non-zero. In general, the singular values of X must be determined by numerical methods. If the ratio of the largest to the smallest singular value of X, that is its condition number $\kappa(X)$, is large enough, then the smallest singular value is indistinguishable from zero relative to

the largest singular value. So, the larger the condition number, the closer a matrix is to having deficient rank.

The SVD provides a characterization of the contrasts of the regression parameter β which are estimable/inestimable. Let $X = UDV^T$, be a SVD, where U is $n \times p$ column orthonormal V is $p \times p$ column orthonormal and D is $p \times p$ diagonal matrix of non-negative singular values. If $V_{[j]}$ is the jth column of V which corresponds to a non-zero singular value of X, then the contrast $\phi = V_{[j]}^T \beta$ is estimable. The linear subspace spanned by all estimable contrasts can be defined by

$$\Phi = \{\phi_l = l^T \beta, \;\; l \in \; row(X)\},$$

where $row(X)$ is the row-space of X which coincides with the $span\{V_{[j]} : D_{jj} \neq 0, j = 1, \ldots, p\}$. Any contrast ϕ_l for which $l \notin row(X)$, is inestimable.

We have just characterized the inestimable contrasts by saying that they are any contrast which is not estimable. A more informative characterization is to note that the subspace orthogonal to $row(X)$, i.e., $row(X)^{\perp}$ consists of $p-$vectors l defining contrasts ϕ_l for which the sample regression provides no information. These are the fundamental quantities which cannot be estimated with the information available in the sample. An operational characterization of these special contrasts is

$$\Phi_0 = \{\phi_l = l^T \beta, \;\; l \in span\{V_{[j]} : D_{jj} = 0, j = 1, \ldots, p\}.$$

Of course, it may be that while the matrix X is rank deficient, the regression (1) is still useful for its estimates of contrasts $\phi_l \in \Phi$. However, we are particularly concerned about those situations when the sole purpose of the analysis is to infer about contrasts which are inestimable or would be inestimable without the presence of a particular case group.

3 Rank-Influential Case Groups: Characterization

For any group $I = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$ of k cases (indices) let $y_{(I)}, X_{(I)}$ denote the reduced data set which excludes those cases in I. A group I is *rank influential* if $X_{(I)}$ is rank deficient, whereas

the complete model matrix X has full rank. We propose identifying rank-influential groups of cases by finding those case groups $I = \{i_1, i_2, \ldots, i_k\}$, for which $|X_{(I)}^T X_{(I)}|$ is zero or close to zero relative to $|X^T X|$. (Of course, if $k > n - p$ then the reduced model matrix $X_{(i)}$ is necessarily rank deficient.)

For any group I, it is well known that

$$|X_{(I)}^T X_{(I)}| = |X^T X| \times |I_k - H_I| \tag{2}$$

where I_k is the order$-k$ identity matrix, and H_I is the $k \times k$ submatrix of $H = X(X^T X)^{-1} X^T$ consisting of those elements H_{ij} where both $i \in I$ and $j \in I$. See, for example, Cook and Weisberg (1982, p. 210).

If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ are the k ordered eigenvalues of H_I, then

$$|I_k - H_I| = \prod_{j=1}^{k} (1 - \lambda_j). \tag{3}$$

It follows that $|X_{(I)}^T X_{(I)}| / |X^T X|$ is zero (or close to zero) if and only if λ_1 is one (or close to one). We shall use $\bar{\lambda}_I$ to denote λ_1, the largest eigenvalue of H_I.

When $k = 1$ or $k = 2$, these eigenvalue computations can be solved easily, albeit separately for each case group. For $k = 1$ and $I = \{i\}$, say, then $\bar{\lambda}_I = H_{ii}$, the ith diagonal element of H. So, rank influence of a case is fully explained by its leverage, as defined by Hoaglin and Welsch (1978). For $k = 2$ and $I = \{i, j\}$, then $\bar{\lambda}_I = [H_{ii} + H_{jj} + \sqrt{(H_{ii} - H_{jj})^2 + 4H_{ij}^2}]/2$. To identify rank-influential pairs directly then, we must consider $\binom{n}{2}$ pairs I and their associated $\bar{\lambda}_I$. To generalize to groups of size $k > 2$, we must solve $\binom{n}{k}$ $k-$order polynomials for their largest roots. The combinatorial and numerical complexities of this approach are clearly non-negligible. (However, some computational savings could be had by exploiting relationships like $I \subset I' \Longrightarrow \bar{\lambda}_I \leq \bar{\lambda}_{I'}$.)

Also, methods based on the singular values or condition number of the reduced matrices $X_{(I)}$ suffer from similar computational burdens. The effects on the singular values of a matrix when k rows are deleted depend on the SVD of H_I as well. Kempthorne (1985) discusses this problem and gives an exact characterization when $k = 1$ and $p = 2$. In

general, the solution requires determining the smallest (and largest) root of a p-th order polynomial for every case group whose rank influence is in question. When $p > 2$ and k is small, Hadi (1988) and Hadi and Wells (1989), propose excellent approximations to changes in the extreme singular values and the condition number when a case group is deleted from the model matrix. However, even with such approximations the combinatorial search over possible subsets of cases remains.

We propose methods in the next section with the express intent of avoiding a search with modest to high numerical and combinatorial complexities. They rely on

Theorem 1. When X is an $n \times p$ model matrix of full rank, a group of cases I is rank influential, that is, $\bar{\lambda}_I = 1$, if and only if there exists a non-null vector $u \in \mathcal{C}$, the column space of X, satisfying $u_i = 0$, for all $i \notin I$.

Proof.

The case group I is rank influential

$\iff$ $X_{(I)}$ is rank deficient

$\iff$ $\exists\, l \in \Re^p$ such that $X_{(I)} l = 0$

$\iff$ $\exists\, l \in \Re^p$ such that $[Xa]_i = 0, \forall\, i \notin I$

$\iff$ $\exists\, u \in \mathcal{C}$ such that $u_i = 0\ \forall\, i \notin I$.

4 Identifying Rank-Influential Groups

By Theorem 1 we can detect rank-influential groups by searching $\mathcal{C}$, the column space of X, for any vectors that are closely aligned to a few coordinate axes and perpendicular to the rest.

The first method exploits factor-analytic rotation algorithms to identify such n-vectors u whose nonzero components identify rank-influential case groups.

4.1 Search by Quartimax/Varimax Rotation

Any unit vector $u' \in \mathcal{C}$ can be expressed as $u' = Ua$ for some unit vector $a \in \Re^P$, where U is any $n \times p$ matrix whose columns form an orthonormal basis for the column space of X (which we assume

has full rank). The matrix Q can be calculated, for example, as the "Q" matrix in the QR-decomposition of X or as the left column-orthonormal matrix in a SVD of of X.

Consider the matrix $U' = UA$, where A is any $p \times p$ orthogonal matrix. The columns of U' also form an orthonormal basis for $\mathcal{C}$. If the first column of A is a, then the first column of U' is u', as before. For any group I, a lower bound for $\bar{\lambda}_I$ can easily be calculated using U', as we now show.

Theorem 2. Suppose U' is an $n \times p$ matrix whose columns form an orthonormal basis for $\mathcal{C}$, the column space of X. Then

$$\bar{\lambda}_I \geq \max_{1 \leq j \leq p} \{\sum_{i \in I} [U'_{ij}]^2\}.$$

Proof. Suppose that the size of group I is k. if the maximum eigenvalue of H_I is $\bar{\lambda}_I$, then, for all k-vectors v, $v^T H_I v / v^T v \leq \bar{\lambda}_I$. Let v be the k-vector with components U_{ij}, $i \in I$ for any fixed j. It follows that $v^T H_I v = \sum_{i \in I} [U'_{ij}]^2$, because $H_I = U_I U_I^T$. Since $v^T H_I v \leq \bar{\lambda}_I$ and j is arbitrary, the conclusion of the theorem follows.

One way to find rank-influential groups then is to transform U orthogonally to U' in such a way that the columns of U' have as many as possible zero or near-zero elements and a few elements with large magnitudes. If there is a case group I and a column j for which $U'_{ij} = 0$, for $i \notin I$, or equivalently $\sum_{i \in I} [U'_{ij}]^2 = 1$, then I is a rank-influential group by Theorems 1 and 2.

Consider transforming U to $U' = UA$ by applying the varimax/quartimax rotation A. Kaiser (1958) proposed the varimax method for transforming a matrix U to U', orthogonally, so as to maximize

$$\psi = \sum_{j=1}^{p} \psi_j,$$

where

$$\psi_j = \frac{1}{n} \sum_{i=1}^{n} [U'_{ij}]^4 - \left(\frac{1}{n} \sum_{i=1}^{n} [U'_{ij}]^2 \right)^2, \quad j = 1, 2, \ldots, p.$$

The term ψ_j is the variance of the squared elements in the jth column of U'. In our case where U is column orthonormal, the sum of the squared elements in each column of U is one, so the varimax criterion is equivalent to choosing the orthogonal transformation to maximize the *quartimax* criterion

$$\psi^* = \sum_{j=1}^{n} \sum_{i=1}^{n} [U'_{ij}]^4.$$

These equivalent rotation criteria were proposed in factor analysis to achieve interpretable factors from column-orthonormal loading matrices. In particular, it is argued that maximization of ψ or ψ^* results potentially in "a factor matrix with a maximum tendency to have both small and large loadings" (Kaiser 1958, p. 188).

We now propose a four-step method for identifying rank-influential groups.

Method 1: Quartimax/Varimax Search of $\mathcal{C}$.

Step 1: Construct U, a column-orthonormal matrix whose columns are a basis for $\mathcal{C}$, the column space of X.

Step 2: Transform U to U', applying the varimax/quartimax rotation.

Step 3: For each column j of U' examine the squared elements and identify those groups I for which $[U'_{ij}]^2 \approx 0$, $i \notin I$ and $[U'_{ij}]^2 > 0$, $i \in I$.

Step 4: For any group I identified in Step 3, calculate

$$\bar{l}_I = \max_{j=1,2,\ldots,p} \sum_{i \in I} [U'_{ij}]^2,$$

the lower bound for $\bar{\lambda}_I$; if $\bar{l}_I = 1$, then classify case group I as rank influential.

Although our interest focuses on case groups I for which $\bar{l}_I = 1$, it is also useful to identify those groups with $\bar{l}_I$ less than one but still large. Draper and John (1981) propose the measure $|I_k - H_I|$ as a multiple-cases extension of the leverage used by J.W. Tukey, Huber (1975), and Hoaglin and Welsch (1978). Our measure $\bar{l}_I$ provides an upper bound for $(1 - \bar{l}_I)$ for $|I_k - H_I|$. Draper and John discuss how

their measure can be interpreted as a factor in the statistic proposed by Andrews and Pregibon (1978) to assess the influence of case groups in linear regression analysis.

Cook and Weisberg (1982, Section 3.6) propose $\bar{\lambda}_I$ and $|I_k - H_I|$ as possible measures of the "potential" of group I in linear regression. When a case group has high potential, a small change in the observed vector y (in certain directions) can result in large changes in the value of the least squares estimate $\hat{\beta} = (X^TX)^{-1}X^Ty$. They use $\bar{\lambda}_I$ to construct upper bounds on the value of a statistic proposed by Cook (1977, 1979) to measure the distance between the least squares estimates of β when a case group is included or excluded from the data. The statistic $\bar{l}_I$ we propose to asess rank influence could also be employed as a diagnostic to screen the data for case groups whose presence in the data may have a large impact on the values of fitted parameters.

We now present three examples. The first is an application to real data which succeeds in identifying a single rank-influential case group. The second uses artificial data to illustrate how the Quartimax/Varimax Search can identify p distinct case groups. The third example also uses simulated data and illustrates how the method can fail.

Example 1. The Florida Area Cumulus Experiment was conducted in 1975 to determine the effect of seeding clouds with silver iodide to increase the amount of rainfall. Cook and Weisberg (1982) use data from this experment to study the detection of influential observations in regression. They give (pp. 3-4) the data together with complete descriptions of each variable. For our purposes we note that y is the response variable, a measure of the amount of rain, and the others are explanatory variables, where the variable "A" indicates seeding (1) or no seeding (0).

The following linear regression model is proposed:

$$\begin{aligned} log_{10}y &= \beta_0 + \beta_1 A + \beta_2 T + \beta_3(S - N_e) + \beta_4 C + \beta_5 log_{10}P + \beta_6 E \\ &\quad + \beta_7[A \times (S - N_e)] + \beta_8[A \times C] + \beta_9[A \times log_{10}P] + \beta_{10}[A \times E] \\ &\quad + \epsilon \end{aligned}$$

The cross-product variables are included to model a possible non-additive treatment effect due to seeding. The X matrix for this model

has 24 rows and 11 columns.

Accepting the legitimacy of fitting such a complex model with only 24 observations, we apply the Quartimax/Varimax Search to identify rank-influential case groups. Table 1 gives U, the left column-orthonormal matrix in the SVD of X, whose columns are a basis for $\mathcal{C}$, the column space of X. Note that U reveals no rank-influential groups; that is no column has nonnegligible elements in only a few rows. The transformed matrix U' was obtained by a varimax rotation of U, and Table 2 gives the squares of the elements in U'. A brief examination of the columns of U' reveals that in the seventh column, all the elements are very close to zero except those in rows 3 and 20. For case group $I = \{3, 20\}$, a simple calculation gives $\bar{l}_I = .999997$. Since the numerical calculations were performed in single precision, we can conclude that $\bar{\lambda}_I = 1$, that is, I is rank influential. Indeed, when cases 3 and 20 are removed from the data, the variables A and $A \times E$ are identical and the X matrix is rank deficient.

The rank influence of this pair was noted by Cook and Weisberg (1982, p. 145), but they calculated $\bar{\lambda}_I$ for all pairs I. They also note that any pair including case 2 has "high potential." Using column 2 of U', we find that $\bar{l}_{\{2\}} = .97283$, confirming that case 2 has high potential. Although $\bar{l}_I$ will exceed $\bar{l}_{\{2\}}$ for any pair I including case 2, the increase is always smaller than 1%. This supports their opinion that case 2 ought not to be viewed as a member of a pair (other than with case 20).

Once a rank-influential case group has been identified, it is useful to understand how the group affects the the estimability of β. The estimability of certain contrasts of β are unaffected. In terms of the SVD of $U_{(I)}$, the contrasts rendered inestimable if case group I is unavailable are $\phi_a = a^T\beta$, for a in the space spanned by the right singular-vectors of U corresponding to unit singular values.

Example 2. To demonstrate that the Quartimax/Varimax Search can identify multiple rank-influential case groups, consider the model matrix associated with a trivial analysis of covariance experiment:

$$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 1 & 3 \\ 0 & 1 & 3 \end{bmatrix}.$$

The first two columns identify one of two treatments and the third column gives the values of a covariate. The regression model consists of two parallel lines describing the relationship between the response and the covariate with the difference between the lines giving the treatment effect. It is easily shown that the least-squares fit to the slope of the parallel lines is inestimable if either $I = \{1,2\}$ or $I' = \{3,4\}$ is excluded. Also, the treatment effect cannot be estimated if $I'' = \{5,6\}$ is excluded.

The left column-orthonormal matrix from the SVD of X is

$$U = \begin{bmatrix} -.2117 & .4282 & -.5214 \\ -.2117 & .4282 & -.5214 \\ -.3844 & .3725 & .4620 \\ -.3844 & .3725 & .4620 \\ -.5544 & -.4218 & -.1213 \\ -.5544 & -.4218 & -.1213 \end{bmatrix}.$$

The three rank-influential groups I, I', I'' are not readily apparent in this basis matrix for $\mathcal{C}$. But, they are highlighted upon applying the Quartimax/Varimax rotation A to U giving

$$U' = UA = \begin{bmatrix} 0 & 0 & -.7071 \\ 0 & 0 & -.7071 \\ 0 & .7071 & 0 \\ 0 & .7071 & 0 \\ -.7071 & 0 & 0 \\ -.7071 & 0 & 0 \end{bmatrix}.$$

While the Quartimax/Varimax Search was successful in examples 1 and 2, it is limited to identifying at most p rank-influential case groups and, multiple groups must lie in orthogonal subspaces. Moreover, there is no guarantee that any rank-influential case groups will

indeed be found. This last shortcoming is demonstrated in

Example 3. Suppose that the column-orthonormal matrix corresponding to a model matrix X is

$$U = \begin{bmatrix} .500 & .289 \\ .500 & .289 \\ 0 & .577 \\ 0 & .577 \\ .7078 & -.409 \end{bmatrix}.$$

The groups $I = \{1, 2, 5\}$, and $I' = \{3, 4, 5\}$, are the minimal rank-influential groups of size no larger than $n - p$=3. (Of course, any subset of size 4 is trivially rank influential.) The identity rotation clearly identifies I as rank influential. Another rotation could yield a U' with a column identifying I' by simply rotating so that a coordinate axis goes through the point (.500, .289). When the matrix U' is constrained to be orthogonal, these two subsets cannot be identified simultaneously.[1]

Allowing oblique rotations might seem an appropriate extension of the orthogonal Quartimax/Varimax Search, but even then, there may be more than p rank-influential groups. While these features clearly limit the potential applicability of the Quartimax/Varimax Search, in this case the varimax rotation of the matrix U yields:

$$U' = \begin{bmatrix} 0.5580 & 0.1490 \\ 0.5580 & 0.1490 \\ 0.1501 & 0.5571 \\ 0.1501 & 0.5571 \\ 0.5763 & -0.5788 \end{bmatrix}.$$

Neither rank-influential subset is identified! Thus, the Quartimax/Varimax Search may fail.

To illustrate the optimization underlying the Quartimax/Varimax search, consider Figure 1. The five cases are plotted in the plane given by the first two columns of the matrix U. The single case is plotted with an asterisk and the case pairs which coincide in the design/model space are represented with xs. The closed curve illustrates the value

[1] I am grateful to a referee for bringing this example to my attention.

of the Quartimax criterion ψ^* for any rotation as the curve's distance from the origin when it intersects any set of orthogonal axes. The axes with intersection furthest from the origin are the quartimax axes.

While the Quartimax/Varimax method may succeed in identifying rank-influential groups for particular examples, it suffers from several potential problems. First, at most p rank-influential groups can be identified. Second, the basis vectors in U' are constrained to be orthogonal. As in example 3, distinct rank-influential groups may lie in oblique subspaces. Third, the Quartimax criterion for specifying the optimal rotation of a basis matrix U may not be appropriate. In the next section, investigation of a component-wise version of the Quartimax criterion will demonstrate that it has a tendency to identify the most outlying case groups in the configuration given by the rows of U. This is evident in Figure 1 for Example 3 – the most outlying case (corresponding to row 5 of U) essentially defines the first Quartimax axis.

4.2 Search by Projection Pursuit

The identification of rank-influential subsets is equivalent to a projection pursuit problem with the n points in p−space whose coordinates are given by the rows of the column-orthonormal matrix U. Consider the problem of finding *interesting* projections of $U : u = Ua$, for some $a \in S^P = \{a \in \Re^p : |a| = 1\}$. Recent applications of projection pursuit have defined *interesting* as those projections achieving a high value of $J(u)$, where the function J is some index based on the empirical distribution of the components of u, for example, the value of the chi-squared statistic for testing normality, the value of the Kolmogorov statistic for testing normality, the modified version of the Friedman-Tukey index, absolute skewness, absolute kurtosis, or Friedman's Legendre polynomial indices; see Huber (1988), also Huber (1985, 1989).

The absolute kurtosis index, $J(u) = \sum_1^n |u_i|^4$, suggests itself naturally in our problem of identifying rank-influential case groups because it corresponds to a one-dimensional or component-wise version of the Quartimax criterion. With an index such as this, we now propose

Method 2: Projection Pursuit Search of $\mathcal{C}$.

Step 1: Specify the Projection Pursuit Index J, for example, $J(u) = \sum_1^n |u_i|^4$.

Step 2: Construct U, a column-orthonormal matrix whose columns are a basis for $\mathcal{C}$, the column space of X.

Step 3: Find the unit $p-$vector a which maximizes $J(Ua)$.

Step 4: Examine the squared elements of $u = Ua$ and identify the group I for which $u_i^2 \approx 0$, $i \notin I$ and $u_i^2 > 0$, $i \in I$.

Step 5: For group I identified in Step 3, calculate

$$\bar{l}_I = \sum_{i \in I} u_i^2,$$

the lower bound for $\bar{\lambda}_I$; if $\bar{l}_I = 1$, then classify case group I as rank influential.

Step 6 (Optional): Repeat Steps 3-5 in the subspace orthogonal to the maximizing a (and orthogonal to any subsequent as).

Huber (1989) details two approaches for identifying maximal directions a in Step 3, depending upon the nature of the index J. For smooth indexes he proposes a variant of the conjugate gradient method and for rough indexes, a random search with similarities to simulated annealing.

Example 3 (Continued). Figure 2 illustrates the results of a Projection Pursuit Search of $\mathcal{C}$ specified by the columns of U in Example 3. We define *interesting* projections of U as those with high values of the component-wise Quartimax (equivalently, absolute Kurtosis) criterion. As in Figure 1, the five cases are plotted in the plane given by the first two columns of the matrix U. The closed curve now illustrates the value of the component wise Quartimax criterion $J(u)$ as the curve's distance from the origin when it intersects any unit vector a emanating from the origin. The Quartimax Axis 1 maximizes the index and Quartimax Axis 2 is a local maximum of the index which happens to be orthogonal to the first. (These axes are identified by the Projection Pursuit Search using Huber's (1988) conjugate gradient search with his equivalent KURT index.)

Our analysis of Example 3 raises the issue of the appropriateness of the component Quartimax criterion for the Projection Pursuit Search (and of the corresponding matrix criterion for the Quartimax/Varimax Search) to identify rank-influential case groups. As noted in the projection pursuit literature, this criterion (which is equivalent to the absolute kurtosis) finds directions exhibiting outliers; see, e.g., Huber (1988, p. 6).

Our search for rank-influential case groups has the complementary purpose – identifying projections in which many cases are inliers, indeed, extreme inliers coincident at the origin. That these two criteria are related in our problem is seen by noting that the orthonormality of the projections of U impose the relationship

$$\sum_{i \in I} u_i^2 = 1 - \sum_{i \notin I} u_i^2.$$

When the case group I is outlying to a maximal extent, the left-hand side of this expression is 1, and the cases $i \notin I$ coincide at the origin.

Apparently, the Quartimax Criterion may overlook those rank-influential case groups which consist of cases which are not outlying in the same direction (e.g., group $I = \{1, 2, 5\}$ in Example 3). However, when they do lie in a similar direction, as in Example 1, then the criterion can succeed.

For general application of the Projection Pursuit Search to succeed, alternative projection indexes are needed. Consider the class of indices

$$J_\alpha(u) = \sum_{i=1}^{n} [1 - |u_i|^\alpha], \quad 0 < \alpha < 1.$$

With values of α close to zero, the index identifies directions with multiple inliers coincident at the origin. In particular,

$$\lim_{\alpha \to 0} J_\alpha(u) = \#\{i : u_i = 0\}.$$

This feature of the indices suggests that we focus on J_α with $\alpha \approx 0$. Define

$$J^*(u) = \lim_{\alpha \to 0} \frac{1}{\alpha} J_\alpha(u) = \sum_{i=1}^{n} (-ln|u_i|).$$

With this new index specifying a Projection Pursuit Search, we now

return to Example 3.

Example 3 (Continued). Figure 3 presents the version of Figures 1 and 2 with the index J^* rather than the matrix/component Quartimax criterion. The axes defined by the extreme values of the J^* index identify all the basic rank-influential groups. While these axes are readily apparent from the figure, they are also identified by the Projection Pursuit Search. We note that modification of Huber's (1988) conjugate-gradient search algorithm was necessary because maxima of the index J^* have singular rather than null derivatives and, local to the maxima, the Hessian is positive definite rather than negative definite.

Group $\{1,2,5\}$ is identified by the horizontal axis, group $\{3,4,5\}$ is found by projecting onto the axis with negative slope and the group $\{1,2,3,4\}$ is determined by the axis with positive slope. Other ("non-basic") rank-influential groups are those which contain one of these three basic groups.

To investigate the effectiveness of the Projection Pursuit Search of $\mathcal{C}$ using the index J^*, consider

Example 4. Five column-orthonormal matrices with $n = 20$ by $p = 2$ were constructed containing single rank-influential case groups ranging in size from $k = 1$ to $k = 16$. Each panels of Figure 5 illustrates the configurations of the 20 cases in two dimensions together with a curve giving the value of J^* in terms of the curve's distance from the origin in projection directions a. While the number of maxima of J^* increases with the size of the rank-influential group k, the index distinguishes rank-influential groups even when their size exceeds 50% of the data. Consequently, we expect this index to prove useful in a wide variety of applications where rank influential groups of even modest size are sought.

5 Summary and Concluding Remarks

While the identification of rank-influential groups of cases is trivial in the case of single-case groups and straightforward with case pairs, the combinatorial and numerical complexity of their direct identification for larger sized groups is considerable. Two alternative methods to combinatorics-based searches are proposed which use searches in $p-$dimensional representations of the n cases of a regression model. The focus of these searches are directions in which most cases are extreme inliers, coincident at the origin.

The first method constructs a column-orthonormal basis matrix for the model matrix using the Quartimax/Varimax rotation of any such basis matrix. While not always successful, the method's computational requirements are modest and it can identify rank-influential groups whose cases are outlying in the design space in similar (that is, other than orthogonal) directions.

The second method applies projection pursuit using a new projection index (J^*) which highlights all rank influential groups, that is, its maximal values occur in projections identifying rank-influential groups. The computations for the projection pursuit search are more involved than the Quartimax-search, but they are less intensive than the alternative combinatorial search for the extreme roots of $p-$th order polynomials where the polynomials are indexed by potentially rank-influential groups.

We conclude with two remarks.

Remark A. The median absolute deviation (MAD) has been suggested as a possible index for the projection pursuit search of degenerate projections.[2] We did not explore use of this criterion for two reasons. First, the gradient of the index at any projection direction depends only on one design point which typically corresponds to just one case or a small number of replicates. In contrast, the gradient of the index J^* depends on all cases. Second, the MAD criterion would lead to the identification of only those rank-influential groups of size smaller than $\frac{n}{2}$. Consequently, it would not succeed in Example 3 where the rank-influential groups constituted 60% of the cases.

Remark B. Rousseeuw's minimum-volume ellipsoid (MVE) has been

[2] Personal communication with A.C. Atkinson.

suggested[3] as potentially useful in identifying rank-influential observations; see Rousseeuw and Leroy (1987, p. 158). When there is a rank-influential group for a given model matrix, no finite-volume ellipsoid based on groups of cases outside the group will contain all the data. The MVE relies upon full-rank configurations of elemental sets of cases. Of course, an ad-hoc method based on the MVE might be explored which perturbs a model matrix by the addition of very small errors to every element. However, if the rank-deficiency corresponds to a direction in which there is little variability, then such "fuzzing up" of the data might hide the rank deficiency.

[3] Personal communication with R.D. Cook.

REFERENCES

Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc., Ser. B.,* **40:** 85-93.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics.* New York: Wiley.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics.* **19:** 15-18.

Cook, R.D. (1979). Influential observations in linear regression. *Technometrics.* **J. Amer. Statist. Assoc. 74:** 169-174.

Cook, R.D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics.* **22:** 495-508.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Draper, N.R. and John, J.A. (1981). Influential observations and outliers in regression. *Technometrics.* **23:** 21-26.

Golub, G.H. (1969). Matrix decompositions and statistical calculations. In *Statistical Computation,* R.C. Milton and J.A. Nelder (Eds.), New york: Academic Press.

Hadi, A.S. (1988). Diagnosing collinearity-influential observations. *Computational Statistics & Data Analysis.* **7:** 143-159.

Hadi, A. S., and Wells, M.T. (1989). Assessing the effects of multiple rows on the condition index of a matrix. Unpublished manuscript.

Harman, H.H. (1967). *Modern Factor Analysis, Second Edition.* Chicago: University of Chicago Press.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.* **32:** 17-22.

Huber, P.J. (1975). Robustness and design. In *A Survey of Statistical Design and Linear Models,* Srivastava, J.N. (Ed.), Amsterdam: North Holland.

Huber, P.J. (1985). Projection Pursuit. Invited discussion paper. *Annals of Statistics.* **13:** 435-525.

Huber, P.J. (1988). An Illustrated Guide to Projection Pursuit. Manuscript and PC-ISP macros.

Huber, P.J. (1989). Algorithms for Projection Pursuit. Manuscript.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika.* **23:** 187-200.

Kempthorne, P.J. (1985). Assessing the influence of single cases on the condition number of a design matrix. Memorandum NS-509, Department of Statistics, Harvard University.

Rousseuw, P.J. and Leroy A.M. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

Stewart, G.W. (1984). Rank degeneracy. *SIAM J. Sci and Statist. Comp.* **5:** 403-413.

Table 1. Q Matrix for the FACE DATA
(x 100000.)

COLUMN

1	2	3	4	5	6	7	8	9	10	11
-782.	17391.	50709.	-15004.	39300.	673.	-14744.	-13944.	3442.	-25163.	-9990.
-3580.	92327.	-14636.	27516.	-13268.	-2137.	-4056.	-3950.	14.	2861.	-1501.
-1935.	10094.	-5123.	-54935.	-12386.	-18991.	-32175.	25410.	-25718.	10854.	10583.
-2220.	6231.	20903.	-14082.	-17823.	-9874.	3330.	5842.	33673.	-1586.	27878.
-3564.	17098.	-6858.	-44396.	-22708.	11645.	22559.	-17784.	-8564.	-11656.	3390.
-4614.	7123.	25872.	-10676.	19924.	-51092.	19038.	-43796.	-27949.	27970.	-14559.
-8635.	2089.	15032.	153.	9537.	-13809.	-524.	-13870.	40383.	19498.	52944.
-11956.	1162.	16372.	-6983.	-28643.	-14603.	16868.	4569.	-1053.	15578.	2709.
-13236.	9787.	41245.	-1437.	16919.	32389.	-2239.	15221.	8347.	6075.	14162.
-13512.	7086.	-8856.	-10283.	16415.	-20650.	31506.	30818.	18053.	-4902.	-17572.
-13986.	4742.	-10290.	-37145.	-12749.	19930.	13319.	-25634.	-7723.	-9980.	7077.
-15213.	604.	-9437.	-21920.	3456.	2204.	6858.	314.	20476.	8664.	-21254.
-15773.	1856.	22035.	-4329.	-30967.	8323.	-3967.	21566.	12769.	-20395.	-2035.
-16611.	372.	-9951.	-17845.	10007.	-4983.	19872.	10382.	14207.	-1983.	-13231.
-18273.	9166.	-11149.	1579.	20540.	-15302.	13190.	24832.	27988.	7954.	-25899.
-18788.	6237.	37057.	-677.	-4526.	26254.	9522.	20000.	-19375.	11055.	-13434.
-24893.	-5923.	9204.	8298.	-15063.	-31720.	-32705.	-18173.	10117.	-60935.	-22593.
-25923.	-3183.	-14231.	-12251.	804.	30761.	-40384.	-35581.	34494.	36474.	-32697.
-26182.	-9133.	1284.	9772.	-36832.	-18242.	1798.	1043.	5554.	101.	4654.
-28035.	3707.	-15900.	-7047.	25632.	-17411.	-47107.	30431.	-22119.	10449.	20739.
-30523.	-6018.	-15810.	-2428.	25163.	1384.	23920.	5090.	-1447.	-17913.	10440.
-31740.	-11383.	14.	20673.	-13530.	-17237.	2003.	-12537.	2524.	20783.	18336.
-38838.	2976.	-21290.	7594.	22149.	21592.	11371.	-21876.	-19493.	-25261.	33191.
-38980.	-8455.	15472.	22464.	-14306.	11687.	6173.	7023.	-33231.	17677.	-18849.

Table 2. Squared Elements of Matrix Q' for FACE DATA
(x 100000.)

COLUMN

1	2	3	4	5	6	7	8	9	10	11
10.	0.	51060.	6.	58.	3063.	0.	6.	1.	3709.	100.
13.	97283.	0.	9.	0.	0.	0.	0.	8.	0.	0.
3816.	65.	10.	8013.	58.	107.	49891.	133.	122.	57.	87.
2670.	6.	272.	2668.	3.	1113.	0.	587.	208.	367.	22637.
409.	393.	15.	38377.	97.	161.	0.	17.	5.	87.	141.
99.	1.	438.	55.	75.	78411.	0.	58.	6.	2.	42.
641.	3.	179.	497.	1594.	317.	0.	221.	47.	948.	51486.
2959.	8.	2288.	2821.	6806.	1714.	0.	729.	223.	2.	2959.
194.	1.	19682.	122.	8148.	2719.	0.	88.	13.	3340.	4225.
91.	8.	3.	24.	15.	31.	0.	2067.	34863.	17.	22.
4730.	106.	3.	26388.	25.	33.	0.	1357.	379.	18.	34.
170.	182.	2.	2278.	12.	17.	0.	6342.	9086.	9.	16.
2440.	5.	414.	2496.	7634.	8062.	0.	509.	193.	5597.	1137.
241.	294.	1.	1573.	9.	14.	0.	168.	14700.	8.	12.
304.	589.	0.	2283.	0.	0.	0.	921.	30990.	0.	0.
317.	1.	5855.	344.	26649.	113.	0.	61.	26.	2109.	967.
56.	0.	465.	31.	80.	0.	0.	32.	4.	76799.	203.
61.	1.	3.	10.	11.	38.	0.	84500.	15.	20.	19.
226.	0.	9684.	261.	6067.	0.	0.	36.	18.	6251.	3578.
3720.	64.	10.	7896.	58.	107.	50108.	132.	121.	57.	87.
18554.	792.	3.	86.	16.	33.	0.	1026.	8335.	18.	25.
1932.	6.	8399.	1657.	4984.	2685.	0.	561.	145.	486.	8965.
55230.	189.	16.	656.	87.	183.	0.	126.	404.	98.	136.
1117.	4.	1199.	950.	37513.	1080.	0.	324.	83.	0.	3120.

Figure 1. Five-Point Representation of U for Example 3.

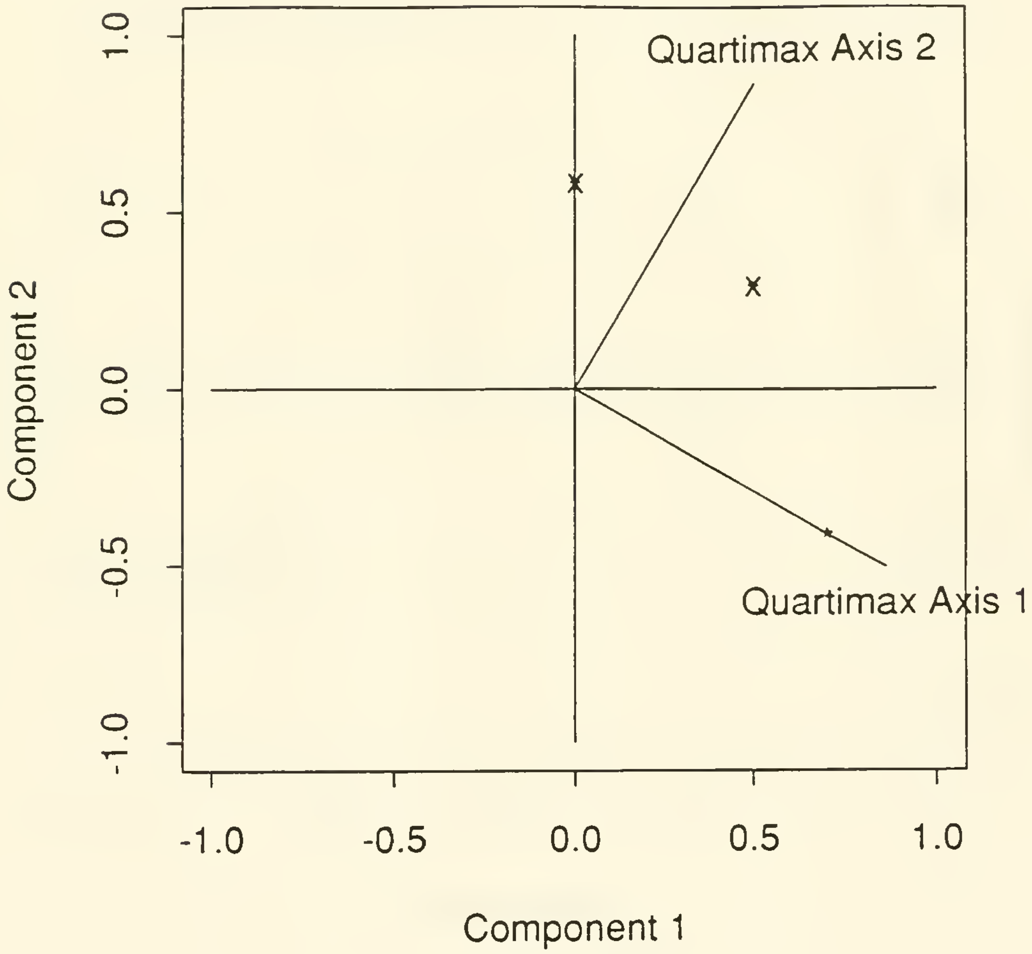


Figure 2. Quartimax Index Plot for Example 3.

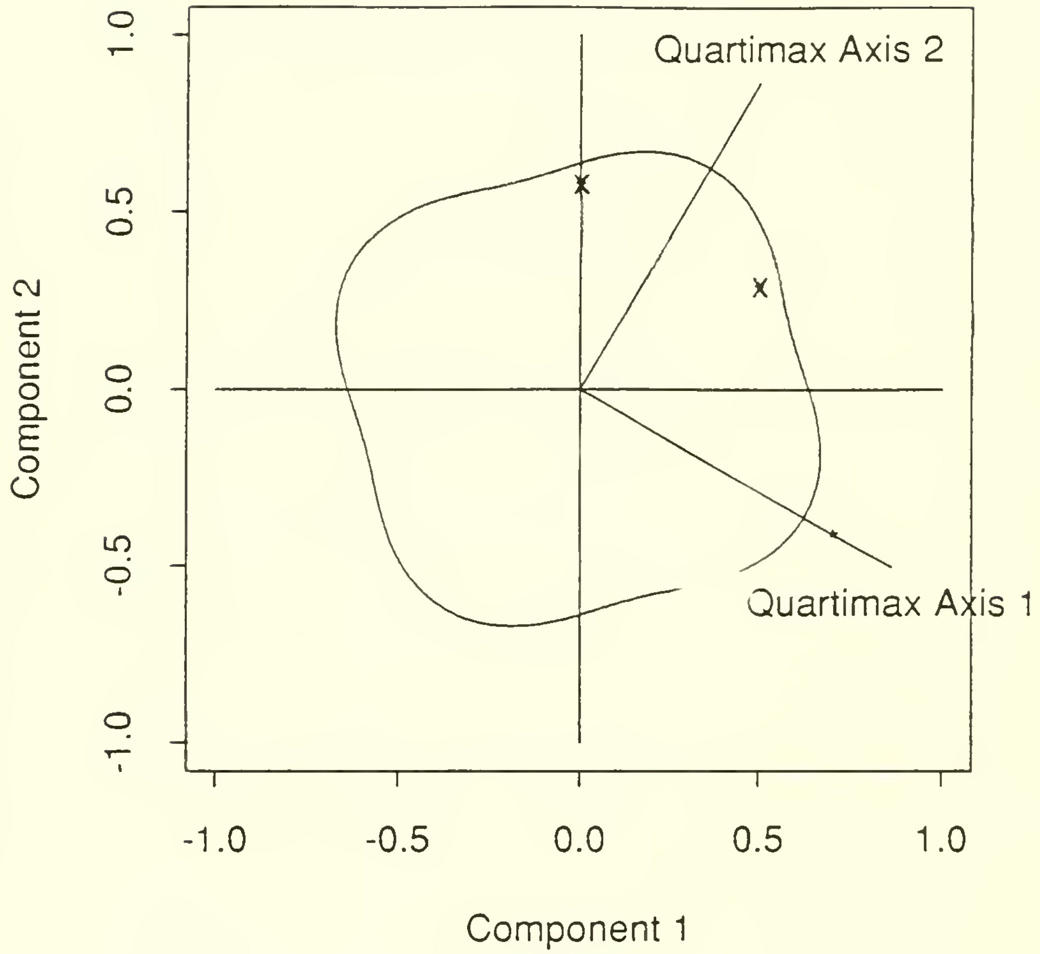


Figure 3. Component Quartimax Index Plot for Example 3.

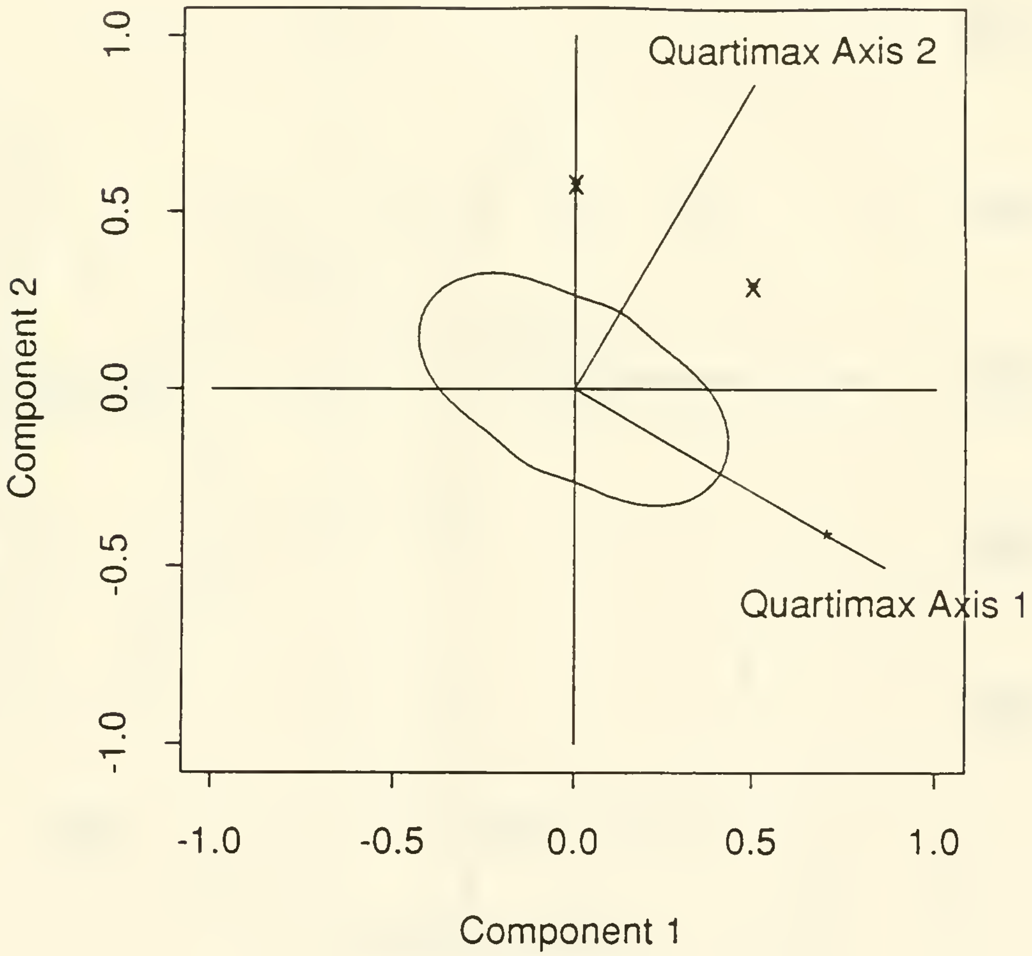


Figure 4. J* Index Plot for Example 3.

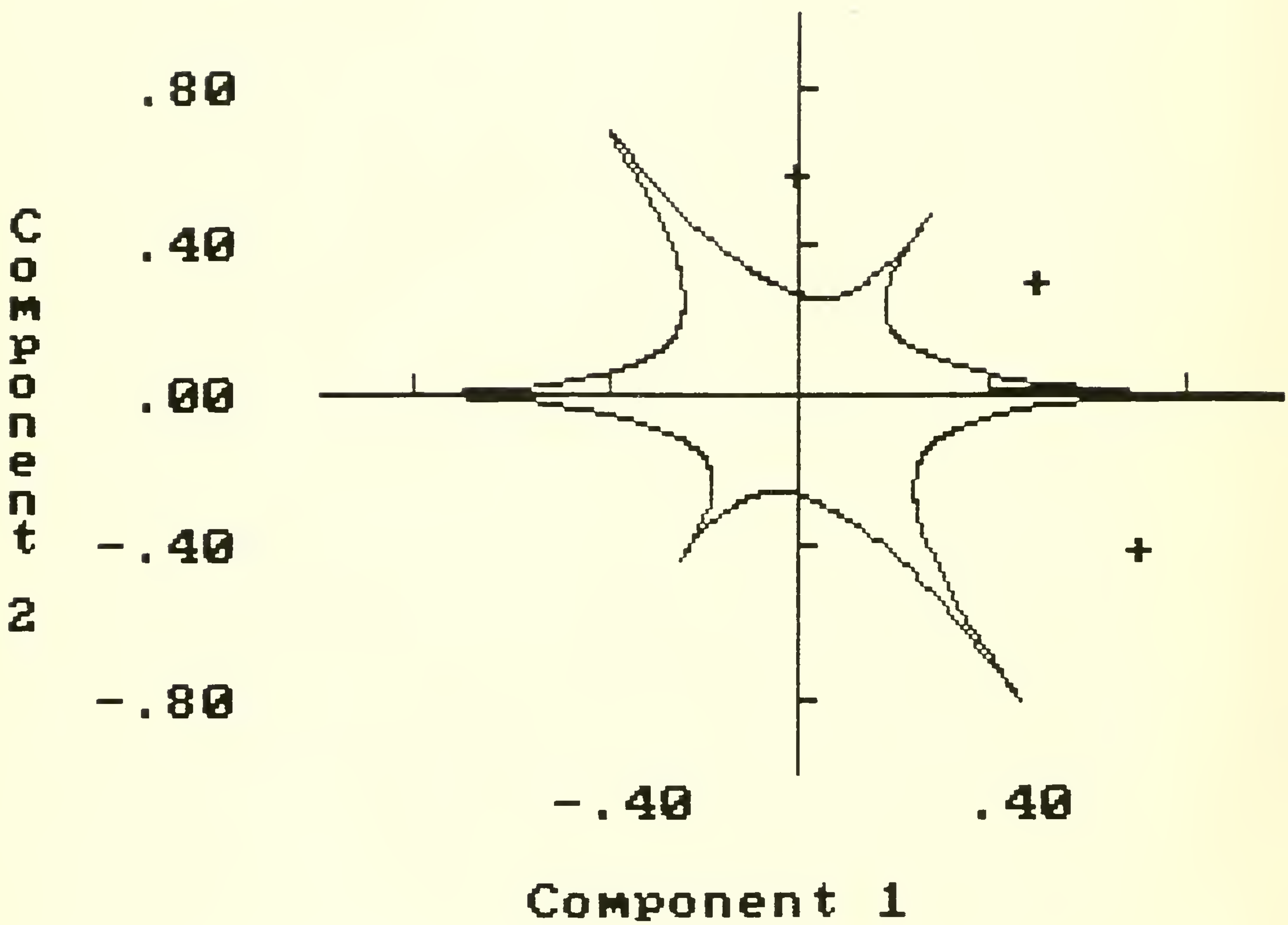
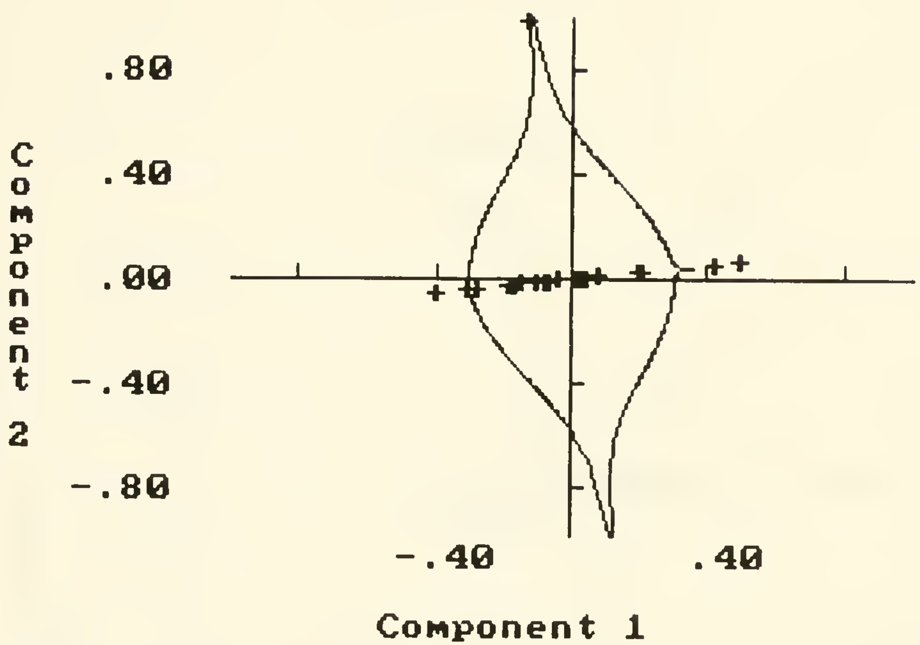
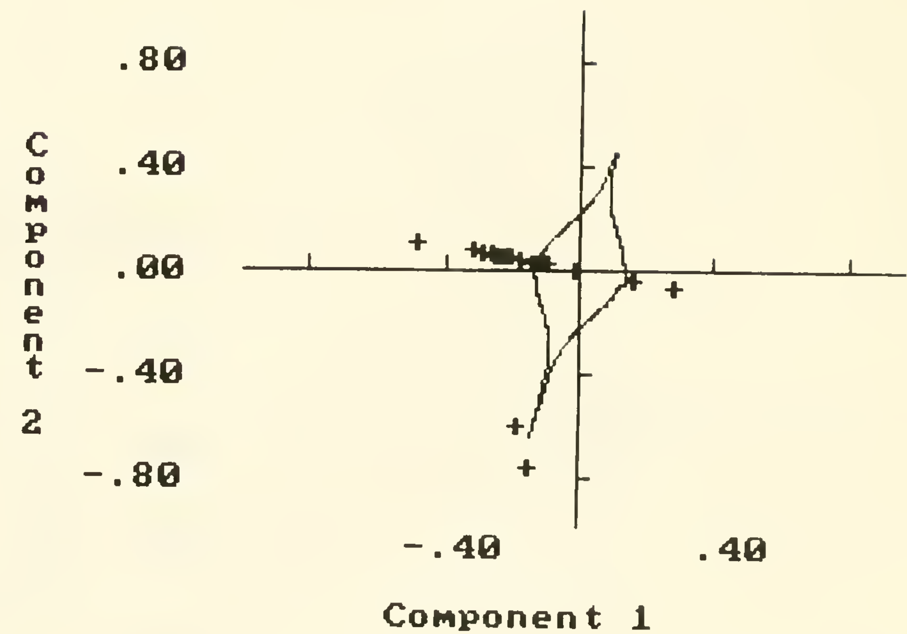


Figure 5.
J* Index Plots for
Five Configurations of n=20 Design Points
(K = Size of Rank Influential Group)

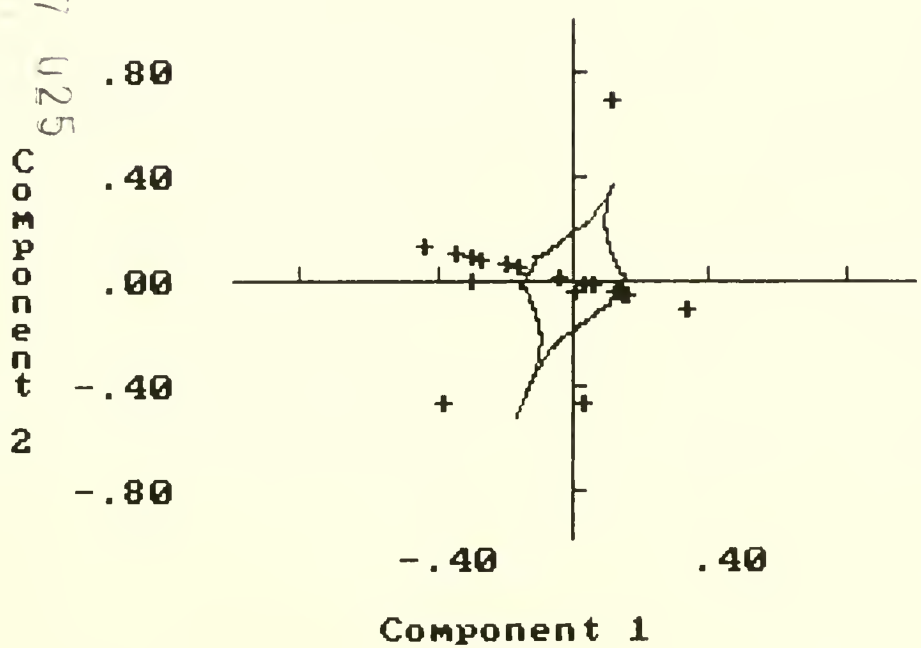


a: K=1, n=20

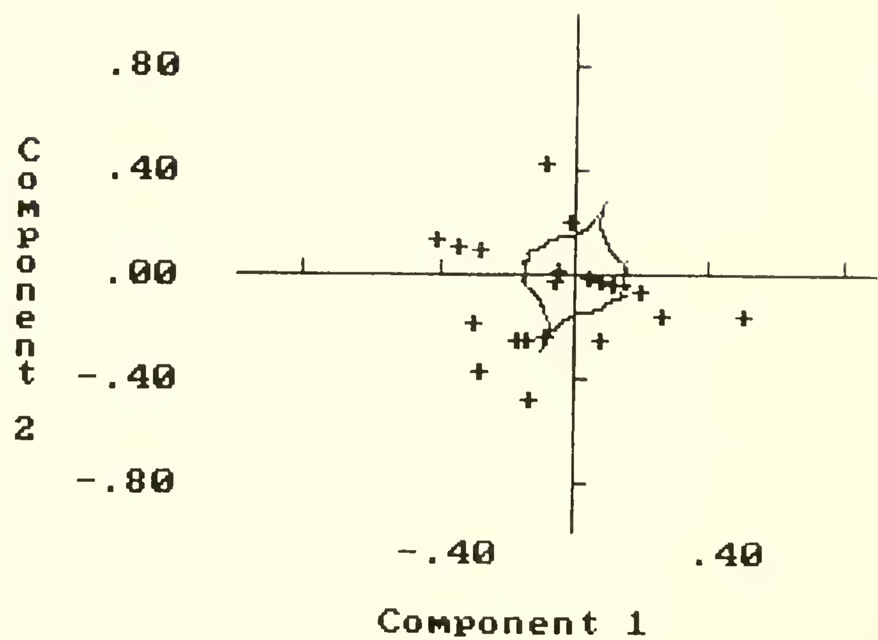


b: K=2, n=20

Figure 5. Continued

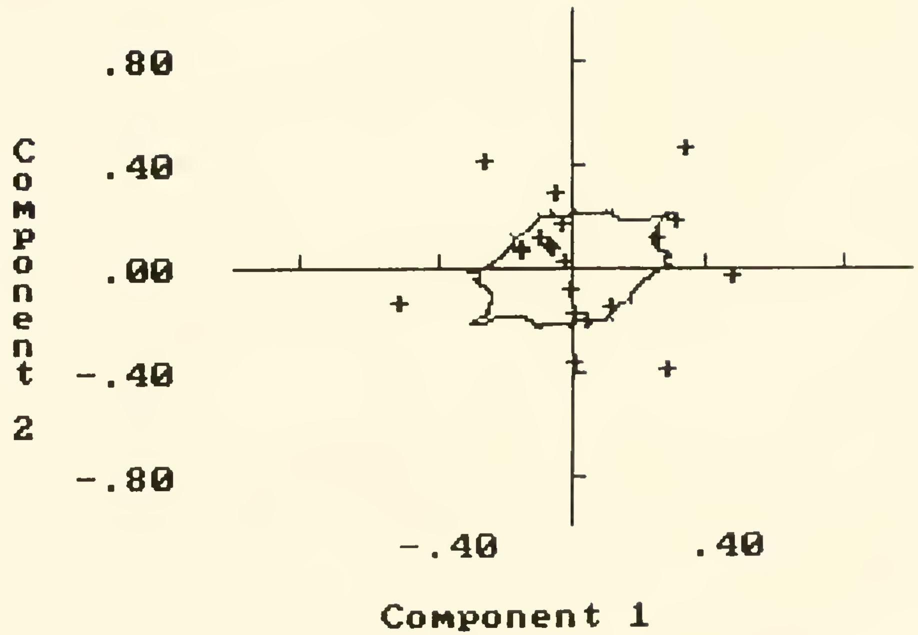


c: K=5, n=20



d: K=12, n=20

Figure 5. Continued



e: K=16, n=20